# 一种知识库体系的设计构建方法及在媒体领域的应用探索

**摘　要：**随着各国政府对知识库的日益重视和大力推动，涌现了大量公共组织探索基于开放数据的知识库构建，具有代表性的如开放链接数据（LOD）项目、在线关联数据知识库 DBPedia 项目等。在企业工程领域，Google、百度、搜狗等也纷纷投身于大规模本体知识库的研究中。随着公共知识库的开放，众多旨在将知识库应用于不同业务领域的领域知识库研究也逐渐开展。本文提出了一种旨在针对媒体应用领域的知识库体系设计构建方法，构建了重点媒体知识库、重点人物知识库、重点事件知识库、业务关键词知识库、业务知识百科库等几大知识库群，详细介绍了知识库构建的几大关键技术，并对知识库在媒体领域的应用场景进行了重点阐述。

**关键词：**知识库；知识图谱；基于本体的知识表示；知识提取；知识标注；知识库应用

**中图分类号：**TP391　　**文献标识码：**A

**文章编号：**1671-0134（2019）05-106-03　　**DOI：**10.19483/j.cnki.11-4653/n.2019.05.035

文 / 陈珺



## 1. 国内外研究现状与发展趋势

随着各国政府对知识库的日益重视和大力推动，涌现了大量公共组织探索基于开放数据的知识库构建。其中，具有代表性的是开放链接数据（LOD）项目，其采用 RDF 形式在 Web 上发布各种开放的数据集，通过来自不同数据源的数据项之间设置 RDF 链接，将不同本体知识库所使用的语义链接互相关联，达到最大程度的全球化知识共享。此外，作为在线关联数据知识库项目，DBPedia 从维基百科的词条中抽取结构化数据，以提供更准确和直接的维基百科搜索，并在其他数据集和维基百科之间创建连接，提供跨语言、跨领域的大规模世界知识。

在企业工程领域，Google、百度、搜狗等也纷纷投身于大规模本体知识库的研究中。比较著名的有“Knowledge Graph”（Google）、“知心”（百度）以及“知立方”（搜狗）等。其通过整合海量的互联网碎片化信息，并将基于围绕关键字的搜索结果知识方式聚合在一起，形成知识集群，对搜索结果进行重新优化计算，将最核心的信息展现给用户。

随着公共知识库的开放，众多旨在将知识库应用在不同业务领域的领域知识库研究也逐渐开展。比如基于案例推理的知识库系统对相关案例的知识进行提取整理，能够为用户输入的问题推荐相似方案与可参考内容。基于本体的专题域知识库系统通过对专题业务资料进行数字化语义处理，并按照本体论思想进行分类标注，实现该业务领域研究的知识集成、知识共享、知识发现和知识重用。

## 2. 知识库总体架构

在媒体领域，相关知识库技术及应用的研究早已开展，并取得了一定的成果。如下是一种知识库体系的总体设计框架，由基础设施层、数据资源层、关键技术层、系统功能层四层组成。




图 1　知识库系统总体架构图

基础环境层主要提供各种所需的计算资源、存储资源、网络资源以及在此基础上搭建起来的大数据基础应用。通过提供关系型数据库、文档知识存储数据库、消息队列和缓存等各种存储形式，实现将不同类型的数据按照其自身特点和业务需求进行分类存储，从而满足系统实时性需求以及系统的分布式响应架构。

数据资源层主要从业务层面提供各种与上层功能相关的各类数据资源的规范存储功能，并提供系统必需的如消息队列、缓存资源等系统数据的统一存储。

关键技术层提供实现系统所需的核心支撑技术系统，提供知识描述与获取、知识图谱、知识库构建与分析研判等关键技术。

功能层主要提供面向业务人员的数据分析与展示功能，以及面向标注人员的人机交互界面。构建重点媒体知识库、重点人物知识库、重点事件知识库、业务关键

词知识库、业务知识百科库五大知识库。每个知识库将实现统一的知识描述方式、分类与组织体系、评价指标体系，最大化兼容现有知识库和功能模块。每个知识库具有知识提取、标注、评估和维护等功能，同时面向标注人员建立评价体系。

基于五大知识库构建相关分析研判功能，包括知识联想与推演、实现业务统计与分析功能。对重要事件进行地域分布、时间周期、人物分布、规律挖掘、趋势预判等分析。

## 3. 知识库构建关键技术

知识库构建关键技术主要包括：知识表示、知识获取、知识图谱、知识持久化和知识评价等几个方面。

### 3.1 知识模式构建

知识表示是知识获取与应用的基础，目前最常用的是基于本体的知识表示方法。本体是对领域实体存在本质的抽象，它强调实体间的关联，并通过多种知识表示元素将这些关联表达和反映出来，这些知识表示元素也被称为原本体，主要包括：（1）概念；（2）属性；（3）关系；（4）函数；（5）公理；（6）实例。总的来说，构造本体的目的是为了实现某种程度的知识共享和重用：（1）本体分析澄清了领域知识的结构，从而为知识表示打好基础。本体可以重用，从而避免重复的领域知识分析。（2）统一的术语和概念使知识共享成为可能。

根据知识来源数据类型的不同，对知识进行分类，形成重要媒体、重要人物、重要事件、业务关键词、业务知识百科五大类别的知识库，每一类知识库可进一步详细分类。采用知识树的方法对知识进行组织，在每一层中，知识节点与其相邻常点在粒度上保持一致。层次越高，粒度越大；层次越低，粒度越小。系统根据用户业务经验针对每一类知识设置对应的树形知识体系。用户可对该体系进行编辑，添加或删除节点，并可对节点名称进行重置。

### 3.2 知识评价体系构建

知识评价体系是对已有知识质量评估的制度，拟从完整度、有效性和相关度三个方面进行评价。知识的完整性由系统根据知识条目属性填充的完整性直接计算得出，完整性计算规则为：权重 * 得分。需与业务人员共同商议确定不同类别知识属性的权重值及分值，基于此给出每一个知识条目完整度百分比。知识的有效性是由系统与业务人员交互得到。业务人员查看某一知识条目时，可以对该知识的有用性进行评价，点击“有用”或“无用”按钮，系统会实时显示每一知识条目的有用性数量分布情况。相关性与有用性相似，由业务人员评价某一知识条目是否与业务相关，指定相关值。若有多人对统一知识条目进行相关性评价时，采用平均值进行显示。

### 3.3 知识获取

知识获取包括：知识提取、知识标注和知识维护等技术。

构建知识库的过程，即是从结构化和非结构化的数据资源中提取知识的过程。结构化知识获取指从特定格式的数据，例如结构化数据库记录、HTML、XML 等含有标签的半结构化数据中进行解析，从而获得多个知识实体及其详细属性，以及知识实体间存在的关联关系。非结构化知识获取指对导入的文本类材料提取文档中提及的实体与关系、要素关键词与文档摘要等，通过自动识别抽取内容的类别将其存储到不同的知识条目集合中。支持的格式包括 TXT、Word、Excel、PDF 等多种形式。

系统支持人工对知识库的词条进行标注与维护，知识标注可采用众包方式。标注人员可通过右键选中某实体对其进行标注，标注的知识在多个数据源中互联互通。若待标注的知识已存在于知识库中，则智能提示补全，节约标注时间，提高标注效率，保证标注的统一性。同时，针对每个知识实体，以可视化的形式对与该实体存在直接关系的关联实体进行展示，并支持对该实体的关联实体及关联关系的可视化编辑。

支持多个用户对知识实体的协同工作，用户修改实体属性后，提交修改时如果数据库中版本与用户修改前版本不一致，系统提醒用户可能产生冲突。用户需获取新版本，并在此基础上进行修改与提交，以保持一致性。

### 3.4 知识图谱

知识图谱将研究知识的关联、联想与推演方法，实现知识推演和研判等应用模式。

知识关联分析对知识库实体间的关联关系进行挖掘与展示，以网络图谱的形式在离散的知识节点之间建立关联关系，当点击关联图中的某一节点时显示关于该节点的详细信息。系统中的知识关联不仅支持同类别的知识实体联系，同样支持重点人物、历史事件、业务关键词、业务知识百科等跨通道知识实体的关联。

知识联想是为了提高业务人员在使用知识库过程中的知识检索效率而提出的。目前，大多数的信息检索采用全文检索技术，检索策略均建立在对于关键词的词频统计学规律上。基于知识联想的检索根据用户搜索内容推荐与该词语义相近的知识条目，为用户提供备选项。

知识推演将根据知识关联图谱中已有的知识，推出新的、未知的知识，以提高知识的完备性，扩大知识的覆盖面，比如同类型知识搜索、关系预测等业务场景。

### 3.5 知识持久化技术

知识持久化技术的目的是将构建出的知识库进行持久化存储。目前，知识图谱中的数据主要采用基于语义的 XML 文档规范、结构化数据库等存储手段进行持久化存储。上述存储手段在进行大规模知识子图查询的过程中，无法在线性时间内实现知识的快速查询。为了加快查询速度，现有查询算法普遍采用图索引技术，但是知识图谱的数据规模大，为其建立图索引需耗费大量的时间和空间开销，从而导致用户难以快速获取满意的查询结果。针对以上特征，我们采用基于图结构（Graph）存储的知识持久化方案，实现快速高效的知识图谱存储与查询。在分布式图数据处理平台的基础上，采用新型的知识图谱查询模型、算法和计算平台分别从知识图谱查询模型、分布式查询算法、分布式查询执行优化三个方

面对知识进行持久化，并提供快速高效的新型分布式查询技术。

## 4. 知识库在媒体领域的应用探索

基于上述构建的重点媒体知识库、重点人物知识库、重点事件知识库、业务关键词知识库、业务知识百科库等几大知识库体系，除能够直接提供相关知识的检索和推荐外，还能够提供知识联想与推演、启发式搜索、个性化推荐、选题深度策划、事件深度分析、趋势预测、机器阅读、机器写作等多种丰富的分析应用功能，可应用于各种新闻生产应用场景。

### 4.1 启发式搜索

对于采编人员或信息分析员而言，很多时候对想要搜索的信息并不是非常确定，因此会先设定一个大致的分析目标，从海量信息中初筛，然后从初筛结果中再调整关键词进一步搜寻更精准的内容，在这个过程中，通过业务领域知识之间的关联关系，可以通过知识联想进行相关知识推荐，从而帮助用户从点到面逐步进行信息的关联分析和深度挖掘，支持这种探索方式的搜索我们可以叫作启发式搜索。比如围绕搜索飞机失事，将相关联信息进行推荐，如飞机失事的历年历史事件追踪、发动机、航空航天、相关制造公司、相关金融股票信息等，通过对大数据基于业务知识关联性的探索，得到更为广泛的分析角度，从而挖掘出更高附加值的信息，加大深度报道产品在社会生活、政治、产业、金融等各个领域的服务价值。

### 4.2 选题策划深度分析

在采编人员针对一个或一组选题进行策划的时候，只推荐出描述上相似的内容很多时候是远远不够的，用户更希望能够挖掘出选题全新的角度，通过从业务领域之间的知识关联上给予知识联想和推荐，这对于一个深度报道、数据新闻和智库咨询的策划是更具有价值的。比如针对雾霾的报道，如果能够超出雾霾本身，挖掘分析雾霾关联的中国能源消耗结构、产业结构和布局，以及拓展到历年国内各项宏观调控政策的影响甚至到海外能源期货大宗市场交易情况等，将会大大提升这类报道内容挖掘分析的广度和深度，提供其他简单同质化报道所不具有的全面性和创新性，从而大大提升媒体报道产品的专业化水平和公众影响力。

### 4.3 事件深度分析

利用媒体行业多维度标签体系，为海量新闻事件进行多维度知识标引，实现具有共指关系新闻内容的专题聚合，基于知识驱动进行各种维度的深入分析，包括事件发生地点、发生时间、事件发生主体、事件相关主体、事件同源关系、事件因果关系、事件时空关系、事件首发媒体、事件涉及的相关政策法规，跟踪事件发展过程中每天的子话题演变过程，并分析国内外重要人物、重要媒体、重要机构关于该事件所发表的观点评述。

趋势预测提供在未来可能发生的重点事件以及可能发生的概率。根据具体业务需求，可包括未来发生事件预测、关键词热度趋势预测、敏感事件信息预测。未来发生事件预测显示未来指定时间段内可能发生的事件及相关信息和发生概率等；关键词热度趋势显示与该事件相关的关键词在指定时间段内的变化趋势；敏感事件预测可提供在未来可能发生的敏感事件及相关信息；用户可自定义时间段来对指定时间段内的趋势进行预测。

### 4.4 机器阅读

机器阅读指用机器自动化完成以前需要人来阅读理解的过程。机器阅读目前比较常见的任务形式是人工合成问答、实体补全和备选答案预测。人工合成问答是经业务人员事先构造好由若干简单事实形成的语料以及相对应的问题，由机器阅读理解文章内容并进行一定的推理，从而得出正确答案；实体补全是在机器阅读并理解语料后，对机器提出相关问题，而问题往往是文章中抽掉实体词的句子，机器回答问题的过程就是预测问题句子中被抽掉的实体词；备选答案预测是机器依据文章、文章的相应问题及候选答案，经过理解和推理，在候选答案中预测出正确答案。通过建立标准化的实体标签，构建知识图谱和领域知识库，能够较好地支持机器阅读上述相关功能的实现。

### 4.5 机器写作

机器写作是一种内容生产的自动化趋势，即基于算法的内容生产和编辑的过程。计算机可根据给定的特定主题，基于特定的算法在已有的备选材料库中选择待组合的内容，通过获取数据、分析数据、提炼观点后以某种特定的格式自动生成内容。

在上述获取数据和分析数据阶段，知识库可以提供机器获取到的特定主题相关的数据以及资料中所提及的与知识库中的知识条目相关的内容信息，用于支撑机器写作过程中的前期数据支撑，同时，能够基于已知的历史知识对其写作结果内容进行丰富。在提炼观点的过程中，知识库的知识条目标签还可以为观点提炼提供基础数据支撑，提高对数据中重要观点的提炼效果。媒

## 参考文献


[1] 陈新蕾，贾岩涛，王元卓，等 . 开发知识库构建技术的多维量化评价方法 [J]. 计算机科学，2017（12）.

[2] 杨玉基，许斌，胡家威，等 . 一种准确而高效的领域知识图谱构建方法 [J]. 软件学报，2018（10）.

[3] 毛麾 . 基于知识库的问答系统 [J]. 现代计算机（专业版），2019（8）.

[3] 官赛萍，靳小龙，贾岩涛，等 . 面向知识图谱的知识推理研究进展 [J]. 软件学报，2018（10）.


（作者单位：新华社技术局）